IRJE (Indonesian Research Journal in Education)

Published by Jambi University, the Graduate School, Doctoral Program in Education

E-ISSN: 2580-5711

Difficulties and Correlation between Phenomenon and Reasoning Tier of Multiple-Choice Questions: A Survey Study

HILMAN QUDRATUDDARSI[1], RENUKA V SATHASIVAM[2], AND HUTKEMRI[3]

Abstract

Two-tier multiple-choice (2TMC) has been introduced as an instrument to analyze a conceptual understanding at tier 1 and reasoning to the answer at tier 2. Even there are many studies on using the instrument; there is still relatively little attention on scoring method. Therefore, this study aimed to compare difficulties and to see if there is a correlation between responses at the phenomenon and reasoning tier. The instrument namely representational systems and chemical reaction diagnostic instrument (RSCRDI) containing 15 items were translated into Indonesian language and validity and reliability of the instrument were established. RSCRDI was tested with 185 pre-service chemistry teachers (19 males and 166 females). Their raw data were converted into logit, and it was found that the phenomenon tier was slightly more difficult. Based on Pearson correlation test, it is found that the phenomenon tier and reasoning tier were significantly correlated, r=.362, p-value <0.001.

Keywords

Chemical reactions, diagnostics test, two-tier multiple-choice questions

1. Postgraduate Student, University Malaya, Kuala Lumpur, Malaysia; hilmandarsi@gmail.com
2. Senior Lecturer, University Malaya, Kuala Lumpur, Malaysia; renukasivam@um.edu.my
3. Senior Lecturer, University Malaya, Kuala Lumpur, Malaysia; hutkemri@um.edu.my

Introduction

Two-tier multiple-choice (2TMC) has been introduced and accepted mainly as an instrument to gauge students' conceptual understanding (Lin, 2016; Saat et al., 2016). The instrument can categorize science students' conceptual understanding into three categories. These categories are the scientific conception, misconception, and lack of knowledge (Akkus, Kadayifci, & Atasoy, 2011; Gurcay & Gulbas, 2015). Discussions on conceptual understanding are widely studied in the literature. The proof is meta-analysis study from Soeharto, Csapo, Sarimanah, Dewi, and Sabri (2019), Gurel et al. (2015) and Wandersee, Mintzes, and Novak (1994) by reviewing 111 articles (2015-2019), 237 articles (1994-2014) and 103 articles (before 1994) on diagnostics instrument. There is also a study from Teo, Goh, and Yeo (2014) who analyzed articles on six best papers on chemistry education-related fields from 2004-2013 and concluded the popularity of conception studies. From the ubiquitous study, 2TMC is recently considered as a sound instrument for which amalgamates the advantages of multiple choices (easy grading, objective) and essay (in-depth analysis)(Gurel, Eryilmaz, & McDermott, 2015; Lin, 2016).

Even there are a plethora of studies on 2TMC in science education, and there is still relatively little attention over how best to analyze two-tier especially grading (Fulmer, Chu, Treagust, & Neumann, 2015). The first study is by Fulmer et al. (2015), who concluded that responding reasoning tier is more demanding compared to answering phenomenon tier. It is in line with studies from Tan, Goh, Chia, and Treagust (2002) who found that around 50.1% of students can correctly answer the phenomenon tier compared to 30.0% in reasoning tier. Therefore, it is recommended to give more points on reasoning tier. To accomplish the effort of findings best way of scoring, Xiao, Han, Koenig, Xiong, and Bao (2018) evaluated six possible scoring systems by measuring psychometric properties such as reliability and goodness of model fit. The differences between six scoring systems are based on the considerations of difficulties between phenomenon and reasoning tier. From the analysis, it is found that there is a scoring system superior to others.

Since there are many possibilities of a scoring system, it is vital to analyze which consideration to take, such as samples' level of education. In the study of Fulmer et al. (2015), there is a different trend in which undergraduate samples can overcome reasoning better compared to phenomenon tier. University students study science concepts are more profound compared to high schools, so it is reasonable if they understand the reasoning of a specific answer. As the proponent of Fulmer et al.'s (2015) study, it is vital to analyze more data for university level, so the consideration of marking system can be addressed correctly. There are two research objectives for this study: (1) to determine the items differ in average estimated difficulties between tier-1 and tier-2 and (2) to determine if any significant correlation of person logit between phenomenon and reasoning tier at 2TMC. There are two research questions concerning this study: (1) How do the items differ in average estimated difficulties between tier-1 and tier-2? And (2) Is there any significant correlation of person logit between phenomenon and reasoning tier at 2TMC

Literature Review

Two-tier multiple choice

2TMC is the extension of TMC by adding more sources of identifying conceptual understanding (Taslidere, 2016). 2TMC has two parts, namely the phenomenon tier and reasoning tier. The first-tier is usually set-up the same way as noted with the ordinary multiple-choice, but the second-tier involves students selecting a reason as to why they chose the answer in the first-tier (Adadan & Savasci, 2012; Schaffer, 2012). Phenomenon tier measures factual knowledge or core concepts in a tested domain (Taber & Tan, 2011). For instance, dilute sulfuric acid is added to some green copper(II) carbonate powder. Vigorous effervescence occurs, and the copper (II) carbonate disappears, producing a blue solution. From the phenomenon, some questions are raised, such as the reasons for the changes, chemical reactions, ionic reactions, and how if reactants are changed (Chandrasegaran, Treagust, & Mocerino, 2007, 2011). The second tier is the justification of responses at phenomenon tier (Taslidere, 2016). In this tier, conceptual knowledge is asked in responses to the phenomenon in the first tier. The rationale of answering phenomenon tier goes beyond knowing (Taber & Tan 2011). Adding reasons can provide profound information about student's conceptual understanding as to the way of assessing learning experience (Fulmer et al. 2015). This reason can be provided as open-ended or multiple-choice as the current study. This method is useful when students provide reasons shortly or insufficient information which tend to be useless and time-consuming (Chu, Treagust, Lim, & Chandrasegaran, 2015).

Figure 1. *Example of two-tier multiple-choice questions*

Phenomenon tier Dilute sulfuric acid is added to some black copper(II) oxide powder and warmed. The copper(II) oxide disappears producing a blue solution. Why is a blue solution produced? A The copper(II) oxide dissolves in the acid producing a blue solution. B Copper(II) oxide reacts with dilute sulfuric acid, producing a soluble salt, copper(II) sulfate. C Copper(II) oxide is anhydrous. When the acid is added the copper(II) oxide becomes hydrated and turns blue.
Reasoning tier The reason for my answer is: 1 The ions in copper(II) sulfate are soluble in water. 2 Cu^{2+} ions have been produced in the chemical reaction. 3 Hydrated salts contain molecules of water of crystallisation. 4 Cu^{2+} ions originally present in insoluble copper(II) oxide are now present in soluble copper(II) sulfate.

Rasch model

The basic principle of this theory is " a person having a greater ability than another person should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult than another means that for any person the probability of solving the second item is the greater one" (Bond & Fox, 2015). In the current study, the Rasch model is utilized to measure the proof of construct validity and data analysis. As stated by Liu (2010) in the book entitled "using and developing measurement instruments in science education: A Rasch modeling approach, the cause of the stagnant replacement of CTT into Rasch model is the lack of training and skills of science educators to apply the theory (Liu, 2010; Romine, Schaffer, & Barrow, 2015).

Data of learning outcomes cannot be treated as interval data because the scoring method by calculating and adding several correct answers can only assume data as ordinal data. Either interval or ordinal data can rank students, but the interval of ordinal data is not equal among data. Therefore the transformation of data is needed to meet the nature of running statistical analysis for comparative study such as t-test, ANOVA, and correlation (Saidfudin et al., 2010). One way to overcome the problem is to transform data by employing the Rasch model. This model can work to address measurement problems by telling the condition when someone responds an item, defining excuses of the responses, directing how to estimate the responses and determining the relation of responses to the estimated situation (Wright, 1977).

In analyzing students learning the outcome, the Rasch model can give a better representation and explanation even in a small number of students. This offers the high precision of comparison and the exact degree of the level of achievement (Osman, Badaruzzaman, & Hamid, 2011). From the Rasch Model, one of exciting feature is also the ability to visualize data using wright map (item-person map) which is a graphical and empirical representation of a progress variable (Boone, Staver, & Yale, 2014; Wilson, 2008).

To estimate respondent measure by considering a person's ability and item difficulties, Rasch model calls the term as logit. Logit = Log (P/(N-P), where P= number of correct item from given items, N= number of given items. Logit is classified into person logit and item logit. Person Logit : Ψ [p] = ln (p/(1-p), item logit : Ψ [p-value] = ln (p-value/(1-p-value), where Ψ symbolize logit transformation.

In nature, the logit score delineates natural log odds of each person to succeed in an item for the determination of the zero point scale (Ludlow & Haley, 1995). Item difficulty is the attribute that affects the person's response while the person's ability shapes the item difficulty estimates (Abdullah, Noranee, & Khamis, 2017). The proponent of Rasch model measurement are two theorems: 1) A more capable person has a higher probability of correctly responding to all the items provided. 2). An easier item is more likely to be answered correctly by all respondents or test-takers (Linacre, 1999; Sumintono & Widhiarso, 2015).

Scoring system in two-tier multiple-choice questions

There are two primary methods of scoring two-tier multiple-choice questions, namely individual scoring and pairing scoring. Pair scoring treats an item pair as a combined item in a dichotomous mode that awards credit for answering both items correctly and zero points for all other responses (Bayrak, 2013; Chandrasegaran et al., 2007; Lin, 2004; Tüysüz, 2009). Individual scoring that treats questions in an item pair as individual items and assigns points for each tier independently (Ding, 2017; Fulmer et al., 2015; Han, 2013; Koenig, Schen, & Bao, 2012; Nieminen, Savinainen, & Viiri, 2012). Based on the study of Fulmer et al. (2015), they found that reasoning tier was more difficult compared to phenomenon tier and suggested to give a higher mark on reasoning tier. Therefore, to extend the study, Xiao *et al.*, (2018) evaluated six methods of scoring as summarized in Table 1. In the table, pattern "00" means incorrect at both phenomenon and reasoning tier, while "11" is correct at both tiers. Pattern "10" is for correct at phenomenon tier only, while "01" means correct at reasoning tier only. By measuring the model fit, it is conceded that no model tend to fit the model better. Therefore, consideration of purposes and sample are essential to determine the best scoring method.

Table 1. *Summary of the scoring method*

Method	Patterns of answer			
	"00"	"10"	"01"	"11"
1	0	0	0	1
2	0	1	1	2
3	0	0	1	2
4	0	1	0	2
5	0	1	2	3
6	0	2	1	3

Reference: (Xiao et al., 2018)

Methodology

Research design

The study was a quantitative study with a survey design. The study stipulates on the collection and explanation of numerical data in the form of answers in phenomenon and reasoning tier (Gay, Mills, & Airasian, 2009; Given, 2008). The main reason for the design is the nature of research objectives and research gaps in the literature. The study is also considered as a survey design because it directly explicates the phenomena of pre-service teacher conceptual understanding without giving any manipulation of the sample characteristics in a point of time (Creswell, 2012).

Instrument

Representational systems and chemical reaction diagnostic instrument (RSCRDI) was adapted from Chandrasegaran et al., (2007) who researched on the development of this instrument. In its application in Singapore, the reliability of the 15-item 2TMC was established by a Cronbach alpha coefficient of 0.65, the difficulty of item ranging from 0.35 to 0.94, and the discrimination indices ranged from 0.35 to 0.59 for 12 of the items (Chandrasegaran, Treagust, & Mocerino, 2009). After obtaining permission to use RSCRDI, the instrument is firstly translated by the first author who studied chemistry in his undergraduate study. The Indonesian language version was reviewed by two chemists who taught for more than ten years and graduated from abroad, specifically the United States and Germany. After the validation, the Indonesian language version is translated back to English by other Chemist who has a PhD in chemistry, and he was graduated in Australia. To establish content validation, the instrument was reviewed by three lecturers who have qualifications looking to their experiences and educational backgrounds. One of them is a professor at inorganic chemistry which suits the materials.

Pilot study

The purpose of this pilot study was to estimate administration time, the reliability, and goodness of model fit of the translated instrument — the sample of the pilot test aged 18-20 years old. They are students from university A as in the real study as many as 69 pre-service chemistry teachers (10 males, 59 females). Analyzing the data of the pilot study applied the scoring rule is dummy variable, 1 for the correct answer, otherwise 0. The result of reliability and separation was in Table 2.

Table 2. *Reliability and separation*

Instrument	Cronbach's Alpha	Person Reliability	Item Reliability	Person separation	Item Separation
Phenomenon	0.59*	0.58*	0.80	1.18	2.01
Reasoning	0.57*	0.58*	0.71	1.04	1.57

Based on Nunnally (1978), the minimum Cronbach's alpha of multiple-choice questions are 0.50, and the value exceeded the minimum score. According to DeVellis (2012), minimally acceptable reliability for person and item are 0.65, and both tiers have person reliability lower than the standard. Reliability score lower than acceptable score are often found in diagnostic test such as 1) (Caleon & Subramaniam, 2010): 0.40 and 0.19 2) (Sreenivasulu & Subramaniam, 2013): 0.40 and 0.43 3) (Sreenivasulu & Subramaniam, 2014): 0.54 and 0.48, 4) (Hoe & Subramaniam, 2016): 0.31 and 0.38, 5) (Yan & Subramaniam, 2018): 0.22 and 0.23. The reliability score for each example is presented for the phenomenon and reasoning tier. It is vital to notice that articles for all example are top-tiered articles (Teo et al., 2014). The next result to consider is item separation, based on (Sumintono &

Widhiarso, 2015), the formula is H (separation) = {(4 x separation) + 1}/3, and the result for phenomenon and reasoning are 3.01 and 2.43, implying that items on phenomenon tier can distinguish test-takers into high, moderate and low, while reasoning tier is only high and low ability.

The following result is the goodness of model fit as in Table 3, which presented mean square (MNSQ), tolerated Z-Standard (ZSTD) and Correlation Points (Pt Mea Corr). Boone, Staver, & Yale (2014) gave the criteria: (a) 0.5 <MNSQ <1,5 (b) -2.0 <ZSTD <+2,0 (c) 0.4 <Pt Measure Right <0.85 or positive measure. To state an item does not fit the Rasch measurement model is the condition in which all criteria fall outside the acceptable value (Sumintono & Widhiarso, 2015). Considering the mean value and item by item, all items fit well the Rasch measurement model. It is concluded that the instrument has good construct validity.

Table 3. *Goodness of model fit*

Item	Infit				Outfit				Pt Mean Corr	
	MNSQ		ZSTD		MNSQ		ZSTD			
	P	R	P	R	P	R	P	R	P	R
1	1.03	0.99	0.2	0	1.13	0.92	0.7	-0.3	0.33	0.34
2	1.1	0.97	1.1	-0.2	1.1	0.93	0.7	-0.4	0.30	0.39
3	0.87	0.96	-0.9	-0.3	0.73	0.93	-1.2	-0.4	0.51	0.4
4	1.03	0.96	0.4	-0.5	0.96	0.96	-0.2	-0.3	0.37	0.42
5	0.95	0.86	-0.4	-1.6	0.88	0.82	-0.5	-1.5	0.43	0.52
6	1.02	0.9	0.3	-0.8	1.27	0.84	1.8	-0.9	0.34	0.49
7	1.02	0.94	0.2	-0.7	1.04	0.92	0.3	-0.6	0.37	0.44
8	1.18	0.95	1.6	-0.5	1.32	0.93	1.7	-0.6	0.19	0.43
9	0.98	1.03	-0.2	0.4	0.91	1.02	-0.6	0.2	0.42	0.36
10	0.86	1.07	-1.6	0.8	0.84	1.05	-1.2	0.5	0.52	0.32
11	0.87	0.9	-1.2	-0.9	0.8	0.82	-1.1	-1	0.51	0.46
12	0.99	1.06	-0.1	0.6	0.96	1.13	-0.2	0.9	0.4	0.30
13	1.2	1.19	1.8	1.9	1.44	1.33	2.2*	2.3*	0.16	0.18
14	0.92	1.07	-0.7	0.7	0.84	1.19	-0.9	1.2	0.47	0.27
15	0.87	1.16	-0.9	1.8	1.01	1.15	0.1	1.3	0.46	0.24
Mean	0.99	1.00	-0.03	0.05	1.02	1.00	0.12	0.03	0.39	0.37
SD	0.11	0.09	0.99	0.99	0.20	0.15	1.12	1.03	0.11	0.10
Min	0.86	0.86	-1.6	-1.6	0.73	0.82	-1.2	-1.5	0.16	0.18
Max	1.2	1.19	1.8	1.9	1.44	1.33	2.2	2.3	0.52	0.52

Data collection and sample

Before data collection, the students were informed that the test was diagnostic, and the results of the test would not affect their grades. However, the result can be beneficial information for their lectures to improve their teaching style (Saat et al., 2016). To collect data, the researcher employs paper and pencil based test which provide a chance for the researcher to observe the process of data collection, better response rate and affordability of respondents (Zuidgeest, Hendriks, Koopman, Spreeuwenberg, & Rademakers, 2011). Since it is a test, data collection is conducted by registering the test class-by-class to selected sample.

In the current study, the selection of the sample is based on stratified random sampling, and the procedure is 1). The population (total 323 pre-service chemistry teachers) was divided into three groups according to their year of education, namely first year, second year and third year. 2). From each group, it was selected 65% (total population, n=209) and randomly selected using SPSS 25. The consideration of choosing the number of samples was the result of calculation employing G* Power 3.0.10.0 where data analysis was one-way ANOVA (effect size 0.40 and alpha value 0.05) was 102 sample. The descriptive statistics of the selected sample are shown in Table 4 below. The sample of the study is 18-21-year-olds with the majority of them from the regency in Lombok and Sumbawa. Only a small number of students are from other provinces such as East Nusa Tenggara. They also from various background of schools such as senior high schools, vocational schools, and Islamic boarding schools.

Table 4. *Demography of the sample of the study*

Characteristics	Number of samples	Percentages (%)
Year of education		
1st year	72	40,19%
2nd year	61	32,06%
3rd year	52	27,75%
Gender		
Male	19	10.27%
Female	166	89.73%
University		
A	114	61.62%
B	51	27.56%
C	20	10.81%
Total	185	100%

Data analysis

To compare the difficulty level of phenomenon and reasoning tier, data of students answer is assumed to measure the same construct. Therefore, reasoning tier of each question are different questions; as a result, there are 30 questions of the study. This methodology was the same as (Fulmer et al., 2015), which carry out a study with similar purposes. The

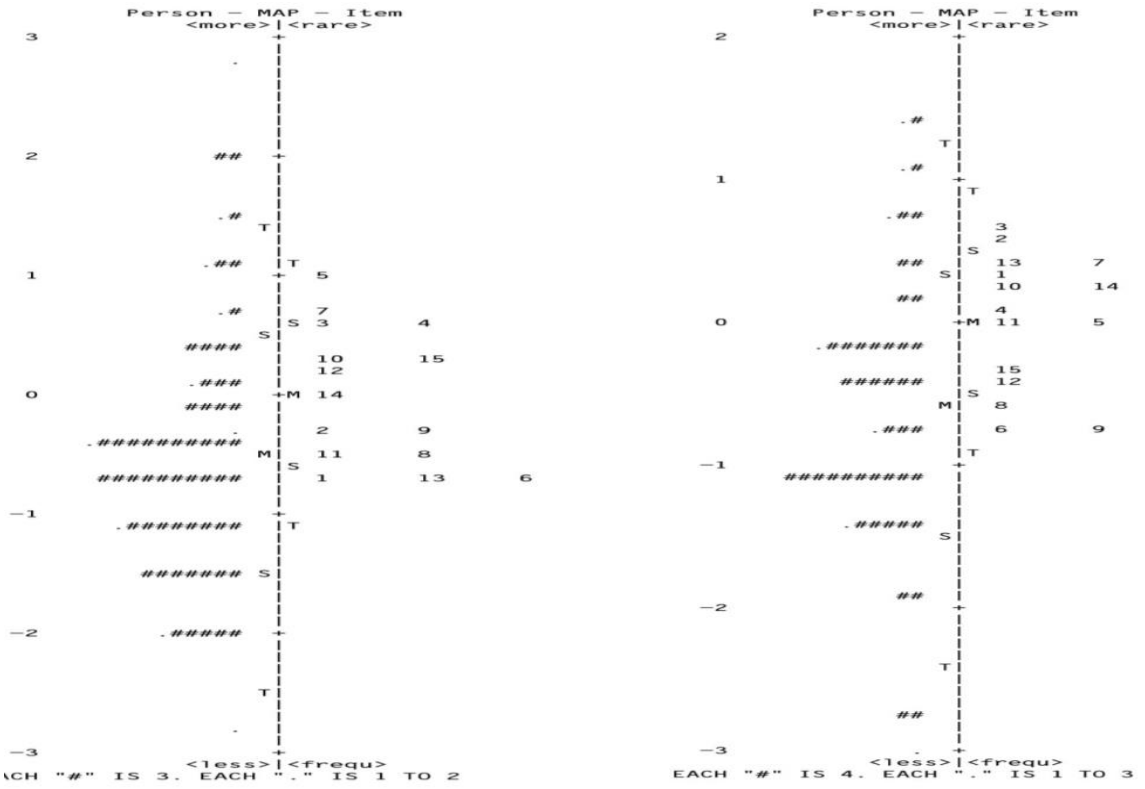
RJE (Indonesian Research Journal in Education)
Published by Jambi University, the Graduate School, Doctoral Program in Education
E-ISSN: 2580-5711

procedures are: 1) Student's responses were coded 1 if they answer correctly and 0 if their answer is incorrect. Then the data were used to find item and student's logit using Winstep 3.7.3 software. Logit is the estimation of respondents' measure by considering a person's ability and item difficulties. 2) Mean of logit from phenomenon and reasoning tier was compared to see their difference. After the comparison, the correlation of both tiers was analyzed to see the tendency of students' responses with the null hypothesis: "student's answer in phenomenon tier is not significantly correlated to student's answer in reasoning tier". The data to use for this analysis is person logit in phenomenon and reasoning tier. From the data, it was estimated normality and linearity as the requisite assumption of the parametric test, then conducted one-tailed Pearson correlation.

Findings

The preview of pre-service chemistry teachers' ability on the chemical reaction on phenomenon and reasoning tier are shown using wright map as in Figure 1. Mean of phenomenon tier is -.626 (SD = 0.945), while reasoning tier is higher with -.525 (0.988). Their mean difference is around 0.1, with almost the same value of standard deviation. It means that the variance of both sets of data looks similar.

Figure 2. *Wright map of phenomenon tier (left) and reasoning tier (right)*



Research Questions 1: Comparison of phenomenon and reasoning tier

To compare the difficulty of phenomenon and reasoning tier, both tiers were assumed to measure the same construct. It implies that in the process of finding logit, each question will be separated into two items to produce 15 items for phenomenon tier and 15 items for reasoning tier. This procedure is adapted from (Fulmer et al., 2015), which also conducted a study on the comparison of components of two-tier multiple-choice questions. From this analysis, it is found the item logit as presented in Table 5.

Table 5. *Logit of each item*

Item	Phenomenon	Reasoning
1	-0.71	0.35
2	-0.33	0.58
3	0.52	0.68
4	0.54	0.08
5	0.91	0.03
6	-0.67	-0.67
7	0.65	0.43
8	-0.54	-0.54
9	-0.35	-0.71
10	0.29	0.24
11	-0.5	0.05
12	0.14	-0.37
13	-0.69	0.43
14	-0.08	0.24
15	0.26	-0.28
Average	-0.037	0.036
SD	0.546	0.452
Min	0.91	-.71
Max	-0.71	0.68

The logit of phenomenon tier is in the range of -.71 to 0.91 with average of -0.037 (SD = 0.546), while reasoning tier has an average of 0.036 in the range of -.71 to 0.68 (SD = 0.452). In the overall, reasoning tier is more complicated than phenomenon tier if referred to average logit, where reasoning is 0.036 > -0.037 (phenomenon tier).

Looking to the data of each item, reasoning tier is more difficult on item 1, item 2, item 3, item 11, item 13, and item 24, while phenomenon tier is more demanding on item 4, item 5, item 7, item 9, item 10, item 12, and item 15, meanwhile 2 items, i.e. item 6 and item 8 have the same level of difficulty. Taking into consideration the average and item comparison, it can be said that their score is almost the same.

Research Questions 2: Correlation of phenomenon and reasoning tier

To see the further relationship of student's performance in the phenomenon and reasoning tier, the Pearson correlation of both scores was conducted to know whether both scores are correlated or not. This analysis is the way of estimating if student answer phenomenon tier correctly, they can have a higher chance to answer the reasoning tier correctly. Before running a correlation test, it is reported the assumption test for correlation to decide the use of the parametric test (Pearson correlation) or non-parametric test (Spearman correlation). These tests are normality and linearity in which normality test is based on the result of skewness and kurtosis, while linearity based on scatter plot of both data sets. Both data are normally distributed, which is indicated by the value of skewness and kurtosis within ±1.96 as the standard of determination normally distributed data. Since both data are normally distributed, the subsequent analysis is to estimate linearity by presenting the result of scatter plots of both data, which assumed that the data fulfil the assumption of linearity.

Since both data are normally distributed, have a linear relationship; fulfill minimum required sample and data scale, one-tailed Pearson correlation can be conducted. The null hypothesis of the analysis is "person logit in the phenomenon tier is not significantly correlated to person logit in the reasoning tier". It is found that phenomenon tier (M=-.6261, SD=.95) and reasoning tier (M=-.5252, SD=.98) is significantly correlated, r=.362, p-value <0.001. It means that a phenomenon tier can explain 13.1% of the variance in reasoning tier. The positive and significant correlation implies that if a student can answer phenomenon tier correctly, there is a high chance of the student to answer the reasoning tier correctly.

Discussion

Difficulties

In this study, it was found that reasoning tier was more difficult than phenomenon tier, meaning that students could express what they know, but they are more difficult to state the reason. The research finding was similar to Fulmer et al. (2015) when analyzing light propagation and visibility data from Chu, Treagust, & Chandrasegaran (2009) which tests 2382 secondary students in Korea and Singapore. Similarly, the study from Liu et al. (2011) also concluded that reasoning tier is more difficult after analyzing the data of 794 middle school students in scientific reasoning. Based on the analysis of using classical test theory (CTT) stated the evidence of the difficulties of requiring students to explain their reasons compared to explicate their knowledge (Caleon & Subramaniam, 2010b; Xiao et al., 2018). As an instance, a study from Tan, Goh, Chia, and Treagust (2002) found that a higher percentage of students can correctly answer phenomenon tier compared to reasoning tier.

From this study, looking to the level of difficulty in each item, more items had greater difficulty on phenomenon tier compared to reasoning tier. This different result is also

revealed by Fulmer et al. (2015), when analyzing the data of the United States on the Classroom Test of Scientific Reasoning (CTSR) with sample undergraduate students. The same thing about this study and CTSR data is the sample from the university. To explicate the conflicting result, it is necessary to look at the teaching style of high school students and university students. In many high schools, specifically Indonesian schools, the study of chemistry covers many topics in a short period. As a result, high schools students study chemistry to fulfill examination, which focuses on the effort to be able to answer a question without understanding the concept in depth. Contrarily, university students learn topics in-depth, which allow them to provide reasoning from chemistry concepts of high schools test.

Correlation

This study found any correlation between the response of students' in phenomenon and reasoning tier. The first possible reason is from the development of the two-tier instrument. In their development, the choices of reasoning tier must be from the common reasons of respondents to select an answer. The careful selection of distractor and key answer is also possibly contributed to the result of correlation.

The other study to find the same result with an identical method is Fulmer et al., (2015). The methodology firstly transforms the data into person logit and measures their correlation. In the study, they found that phenomenon and reasoning tier is correlated for light propagation and visibility in the first data set. In the second data set, both tiers are correlated for control of variables, combinatorial reasoning, probabilistic reasoning and proportional reasoning.

Another study found a correlation between the phenomenon and reasoning tier is Liu et al. (2011) on the topic of energy concepts. In the study, the instrument has ten items with two different forms, namely constructed responses (CR) and explanation multiple-choice questions (EMC). CR items allow students to give reasoning like open-ended 2TMC, while EMC looks like 2TMC, but it is constructed like ordered multiple-choice questions. In the study, phenomenon and reasoning tier for all CR items are significantly correlated, while EMC item has nine items correlated significantly.

Conclusion and recommendations

This study found that reasoning tier is more difficult than phenomenon tier after considering their average, but the comparison of each item shows that they have a fair degree of difficulties. Therefore, this finding suggests considering the level of education of sample (high schools or undergraduate) when an educator wants to score students in responding to two-tier multiple-choice questions, which means that university students do not need any difference in scoring between phenomenon and reasoning tier. Also, the answer of students in both phenomenon and reasoning tier are positively correlated. It means that if students answer correctly in phenomenon tier, there is a high chance of the students to answer reasoning tier correctly.

Disclaimers

This study was funded by Lembaga Pengelola Dana Pendidikan (LPDP) as part of the first author two-year full scholarship to study his master degree.

References

Abdullah, N., Noranee, S., & Khamis, M. R. (2017). The use of rasch wright map in assessing conceptual understanding of electricity. *Social Sciences & Humanities*, *25*, 81–92.

Adadan, E., & Savasci, F. (2012). An analysis of 16-17-year-old students' understanding of solution chemistry concepts using a two-tier diagnostic instrument. *International Journal of Science Education*, *34*(4), 513–544.

Akkus, H., Kadayifci, H., & Atasoy, B. (2011). Development and application of a two-tier diagnostic test to assess secondary students' understanding of chemical equilibrium concepts. *Journal of Baltic Science Education*, *10*(3), 146–155.

Bayrak, B. K. (2013). Using Two-Tier test to identify primary students ' conceptual understanding and alternative conceptions in acid base. *Mevlana International Journal of Education (MIJE)*, *3*(2), 19–26.

Bond, T. ., & Fox, C. M. (2015). *Applying the Rasch model fundamental measurement in the human sciences* (3rd ed.). New York, NY, US: Routledge Taylor & Francis Group.

Boone, W. ., Staver, J. ., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Dordrecht: Springer Netherlands.

Caleon, I. S., & Subramaniam, R. (2010a). Do students know What they know and what they don't know? Using a four-tier diagnostic test to assess the nature of students' alternative conceptions. *Research in Science Education*, *40*(3), 313–337.

Caleon, I., & Subramaniam, R. (2010b). Development and application of a three-tier diagnostic test to assess secondary students' understanding of waves. *International Journal of Science Education*, *32*(7), 939–961.

Chandrasegaran, A. L., Treagust, D. F., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, *8*(3), 293.

Chandrasegaran, A. L., Treagust, D. F., & Mocerino, M. (2009). Emphasizing multiple levels of representation to enhance students' understanding of the changes occuring during chemical reaction. *Journal of Chemical Education*, *86*(12), 1433–1436.

Chandrasegaran, A. L., Treagust, D. F., & Mocerino, M. (2011). Facilitating high school students' use of multiple representations to describe and explain simple chemical reactions. *Teaching Science*, *57*(4), 13–19.

Chu, H.-E., Treagust, D. F., & Lim, G. S. E. Chandrasegaran, A. L. (2015). Efficacy of multiple choice items: Do two-tier multiple-choice diagnostic items have the power to measure students' conceptions similar to open-ended items? *Paper Presented at the*

11th European Science Education Research Association (ESERA) Conference, Helsinki, Finland.

Chu, H., Treagust, D. F., & Chandrasegaran, A. L. (2009). A stratified study of students' understanding of basic optics concepts in different contexts using two-tier multiple-choice items. *Research in Science & Technological Education*, *27*(3), 253–265.

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Boston, MA: Pearson.

DeVellis, R. F. (2012). *Scale development: Theory and applications*. Newbury Park, CA: Sage Publications.

Ding, L. (2017). Progression trend of scientific reasoning from elementary school to university : a large-scale cross-grade survey among chinese students. *International Journal of Science and Mathematics Education*. https://doi.org/10.1007/s10763-017-9844-0

Fulmer, G. W., Chu, H.-E., Treagust, D. F., & Neumann, K. (2015). Is it harder to know or to reason? Analyzing two-tier science assessment items using the Rasch measurement model. *Asia-Pacific Science Education*, *1*(1), 1.

Gay, L. R., Mills, G. E., & Airasian, P. W. (2009). *Educational research: Competencies for analysis and applications* (9th ed.). Upper Saddle River, New Jersey: Prentice Hall.

Given, L. M. (2008). *The Sage encyclopedia of qualitative research methods*. Los Angeles, Calif.: Sage Publications.

Gurcay, D., & Gulbas, E. (2015). Development of three-tier heat, temperature and internal energy diagnostic test. *Research in Science & Technological Education*, 1–21.

Gurel, D. K., Eryilmaz, A., & McDermott, L. C. (2015). A review and comparison of diagnostic instruments to identify students' misconceptions in science. *Eurasia Journal of Mathematics, Science and Technology Education*, *11*(5), 989–1008.

Han, J. (2013). Scientific reasoning: Research, development, and assessment. *PhD Dissertation Ohio State University*.

Hoe, K. Y., & Subramaniam, R. (2016). On the prevalence of alternative conceptions on acid-base chemistry among secondary students: Insights from cognitive and confidence measures. *Chemistry Education Research and Practice*, *17*(2), 263–282. https://doi.org/10.1039/c5rp00146c

Koenig, K., Schen, M., & Bao, L. (2012). Explicitly targeting pre-service teacher scientific reasoning abilities and understanding of nature of science through an introductory science course. *Science Educator*, *21*(2), 1–9.

Lin, J. W. (2016). Development and evaluation of the diagnostic power for a computer-based two-tier assessment. *Journal of Science Education and Technology*, *25*(3), 497–511.

Lin, S.-W. (2004). Development and application of a two-tier diagnostic test for high school students' understanding of flowering plant growth and development. *International Journal of Science and Mathematics Education*, *2*(2), 175–199.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, *3*(2), 103-122.

Liu, O. L., Lee, H., Linn, M. C., & Liu, O. L. (2011). An investigation of explanation multiple-choice items in science assessment. *Educational Assessment*, *16*(3), 164–184.

Liu, X. (2010). *Using and developing measurement instruments in science education: A Rasch modeling approach*. Charlotte, NC: Information Age.

Ludlow, L. H., & Haley, S. M. (1995). Rasch model logits: interpretation, use and transformation. *Educational and Psychological Measurement*, *55*(6), 967–975.

Nieminen, P., Savinainen, A., & Viiri, J. (2012). Relations between representational consistency , conceptual understanding of the force concept , and scientific reasoning. *Physical Review Special Topics- Physics Education Research*, *010123*(8), 1–10.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Osman, S. A., Badaruzzaman, W. H. W., & Hamid, R. (2011). Assessment on students performance using rasch model in reinforced concrete design course examination. *Recent Researches in Education*, 193–198.

Romine, W. L., Schaffer, D. L., & Barrow, L. (2015). Development and application of a novel rasch-based methodology for evaluating multi-tiered assessment instruments: validation and utilization of an undergraduate diagnostic test of the water cycle. *International Journal of Science Education*, *37*(16), 2740–2768.

Saat, R. M., Hidayah, M. F., Aziz, N. A. A., Haron, K., Rashid, K. A., & Shamsuar, N. R. (2016). Development of an online three-tier diagnostic test to assess pre-university students ' understanding. *Journal of Baltic Science Education*, *15*(4), 532–546.

Saidfudin, M., Azrilah, A., Rodzo'an, N., Omar, M., Zaharim, a, & Basri, H. (2010). Easier learning outcomes analysis using Rasch model in engineering education research. *Latest Trends on Engineering Education*, 442–447.

Schaffer, D. (2012). An analysis of science concept inventories and diagnostic tests: Commonalities and differences. *Annual International Conference of the National Association for Research in Science Teaching*, (April).

Soeharto, S., Csapo, B., Sarimanah, E., Dewi, F. I., & Sabri, T. (2019). A review of students ' common misconceptions in science and their diagnostic assessment tools. *Jurnal Pendidikan IPA Indonesia*, *8*(2), 247–266. https://doi.org/10.15294/jpii.v8i2.18649

Sreenivasulu, B., & Subramaniam, R. (2013). University students' understanding of chemical thermodynamics. *International Journal of Science Education*, *35*(4), 601–635.

Sreenivasulu, B., & Subramaniam, R. (2014). Exploring undergraduates' understanding of transition metals chemistry with the use of cognitive and confidence measures. *Research in Science Education*, *44*(6), 801–828.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan Rasch pada assessment pendidikan*. Cimahi: Trim Komunikata.

Taber, K. S., & Tan, K. C. D. (2011). The insidious nature of 'hard-core' alternative conceptions: Implications for the constructivist research programme of patterns in high school students' and pre-service teachers' thinking about ionisation energy. *International Journal of Science Education*, *33*, 259–297.

Tan, K. C. D., Goh, N. K., Chia, L. S., & Treagust, D. F. (2002). Development and application of a two-tier multiple-choice diagnostic instrument to assess high school

students' understanding of inorganic chemistry qualitative analysis. *Journal of Research in Science Teaching*, *39*(4), 283–301.

Taslidere, E. (2016). Development and use of a three-tier diagnostic test to assess high school students ' misconceptions about the photoelectric effect. *Research in Science & Technological Education*. https://doi.org/10.1080/02635143.2015.1124409

Teo, T. W., Goh, M. T., & Yeo, L. W. (2014). Chemistry education research trends: 2004-2013. *Chemistry Education Research and Practice*, *15*, 470–487.

Tüysüz, C. (2009). Development of two-tier diagnostic instrument and assess students' understanding in chemistry. *Scientific Research and Essay*, *4*(6), 626–631.

Wandersee, J. H., Mintzes, J. J., & Novak, J. D. (1994). Research on alternative conceptions in science. *In D. L. Gabel (Ed.), Handbook of Research on Science Teaching and Learning (Pp.177-210). New York: Macmillan.*

Wilson, M. (2008). Cognitive diagnosis using item response models. *Zeitschrift Für Psychologie / Journal of Psychology*, *216*(2), 74–88. https://doi.org/10.1027/0044-3409.216.2.74

Wright, B. D. (1977). Solving Measurement Problems with the Rasch Model. *Journal of Educational Measurement*, *14*(2), 97–116.

Xiao, Y., Han, J., Koenig, K., Xiong, J., & Bao, L. (2018). Multilevel Rasch modeling of two-tier multiple choice test: A case study using Lawson's classroom test of scientific reasoning. *Physical Review Physics Education Research*, *14*(2), 020104.

Yan, Y. K., & Subramaniam, R. (2018). Using a multi-tier diagnostic test to explore the nature of students' alternative conceptions on reaction. *Chemistry Education Research and Practice*, *19*, 213–226. https://doi.org/10.1039/c7rp00143f

Zuidgeest, M., Hendriks, M., Koopman, L., Spreeuwenberg, P., & Rademakers, J. (2011). Comparison of a postal survey and mixed-mode survey using a questionnaire on patients' experiences with breast care. *Journal of Medical Internet Research*, *13*(3).

Biographical notes

HILMAN QUDRATUDDARSI is a postgraduate student, University Malaya, Kuala Lumpur, Malaysia, e-mail; hilmandarsi@gmail.com

RENUKA V SATHASIVAM is a senior lecturers, University Malaya, Kuala Lumpur, Malaysia, e-mail; renukasivam@um.edu.my

HUTKEMRI is a senior lecturer, University Malaya, Kuala Lumpur, Malaysia, e-mail: *hutkemri*@um.edu.my